Biophysical Journal Volume 70 March 1996 1183–1197

Modeling of Polypeptide Chains as C_α Chains, C_α Chains with C_β, and C_α Chains with Ellipsoidal Lateral Chains

Federico Fogolari,* Gennaro Esposito,* Paolo Viglino,* and Serge Cattarinussi‡
*Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, 33100 Udine, Italy, and ‡Institut de Chimie Physique, Departement de Chimie, École Polytechnique Fédérale Lausanne, CH-1015 Lausanne, Switzerland

ABSTRACT In an effort to reduce the number of degrees of freedom necessary to describe a polypeptide chain we analyze the statistical behavior of polypeptide chains when represented as C_α chains, C_α chains with C_β atoms attached, and C_α chains with rotational ellipsoids as models of side chains. A statistical analysis on a restricted data set of 75 unrelated protein structures is performed. The comparison of the database distributions with those obtained by model calculation on very short polypeptide stretches allows the dissection of local versus nonlocal features of the distributions. The database distribution of the bend angles of polypeptide chains of pseudo bonded C_α atoms spans a restricted range of values and shows a bimodal structure. On the other hand, the torsion angles of the C_α chain may assume almost all possible values. The distribution is bimodal, but with a much broader probability distribution than for bend angles. The C_α–C_β vectors may be taken as representative of the orientation of the lateral chain, as the direction of the bond is close to the direction of the vector joining C_α to the ad hoc defined center of the "steric mass" of the side chain. Interestingly, both the bend angle defined by $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$ and the torsional angle offset of the pseudo-dihedral $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\beta i+2}$ with respect to $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ span a limited range of values. The latter results show that it is possible to give a more realistic representation of polypeptide chains without introducing additional degrees of freedom, i.e., by just adding to the C_α chain a C_β with given side-chain properties. However, a more realistic description of side chains may be attained by modeling side chains as rotational ellipsoids that have roughly the same orientation and steric hindrance. To this end, we define the steric mass of an atom as proportional to its van der Waals volume and we calculate the side-chain inertia ellipsoid assuming that the steric mass of each atom is uniformly distributed within its van der Waals volume. Finally, we define the rotational ellipsoid representing the side chain as the uniform density ellipsoid possessing the same rotationally averaged inertia tensor of the side chain. The statistics of ellipsoid parameters support the possibility of representing a side chain via an ellipsoid, independently of the local conformation. To make this description useful for molecular modeling we describe ellipsoid-ellipsoid interactions via a Lennard-Jones potential that preserves the repulsive core of the interacting ellipsoids and takes into account their mutual orientation. Tests are performed for two different forms of the interaction potential on a set of high-resolution protein structures. Results are encouraging, in view of the drastic simplifications that were introduced.

INTRODUCTION

The determination of the conformation (or possibly the conformations) that a biomolecule can assume, consistently with its given chemical structure, is one, if not the main, goal of structural biology. The unrestricted search for low-energy regions in the phase space of the molecular degrees of freedom invariably requires unaffordable computer time. The combinatorial nature of the problem is always disguised by imposing restraints or constraints on the possible conformations. These restrictions are derived from experimental data (e.g., from nuclear magnetic resonance (NMR) or x-ray structural determination) or are based on prior knowledge (e.g., on homology modeling, biased build-up procedure, lattice models, etc.) (Abagyan and Totrov, 1994; Eisenmenger et al., 1993; Hunt et al., 1994; Scheraga, 1993; and references on the previous works of his group cited therein).

All of the approaches that perform conformational searches not subject to experimental restraints must reduce the number of degrees of freedom or restrict the conformational space available for each degree of freedom.

To simplify the three-dimensional representation of molecules, C_α carbon coordinates (sometimes embedded in a regular lattice), rather than the complete set of coordinates for the corresponding residues, have often been employed (e.g., Levitt and Warshel, 1975; Levitt, 1976; Godzik et al., 1993). Furthermore, pseudopotentials describing the contact interaction among any pair of amino acids have been proposed based on steric hindrance of residues, observed frequencies of contacts, and/or structural propensities observed in selected sets of protein structures representative of the whole structural database (Miyazawa and Jernigan, 1985; Hinds and Levitt, 1992, 1994; Brower et al., 1993; Goldstein et al., 1992; Gerber, 1992; Kolinski and Skolnick, 1994a).

After the degrees of freedom of the polypeptide have been drastically reduced, specific algorithms are devised to search the conformational space in the most efficient way. With few exceptions (Levitt, 1976; Rackovski, 1990;

Received for publication 12 June 1995 and in final form 30 November 1995.

Address reprint requests to Dr. Federico Fogolari, Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, Via Gervasutta 48, 33100 Udine, Italy. Tel.: 39-432-523085; Fax: 39-432-600828; E-mail: biofisica@dstb.uniud.it.

© 1996 by the Biophysical Society

0006-3495/96/03/1183/15 $2.00

Godzik et al., 1993; Oldfield and Hubbard, 1994; De Witte and Shakhnovich, 1994) little attention has been paid to the first step of this procedure, i.e., the "schematization" of the polypeptide chain. Possible reasons for this are i) the model is so "coarse grained" that spatial details regarding amino acids are considered immaterial; ii) including other features would prohibitively increase the computational load; iii) in the field of statistical mechanics none of the statistical properties of a (very large) ensemble of particles depend on the lattice employed to model the system. Although all of these motivations appear reasonable, it must be considered that lattice models in statistical physics are used to derive properties of a class of molecules, rather than a single molecule, under the assumption that the system has very large dimensions, which is contrary to the case of a biomolecule. In this context, the most correct use of lattice models is that aimed at deriving general properties of proteins rather than the conformation of a single protein (Cattarinussi and Jug, 1990, 1991; Abkevich et al., 1994; Šali et al., 1994; Skolnick and Kolinski, 1990a; Sikorski and Skolnick, 1990; Wolynes et al., 1995), although lattice models have also been used to predict protein tertiary structure with remarkable success (Skolnick and Kolinski, 1990b; Kolinski and Skolnick, 1994b). Another point against excessive simplification is that the computational load from considering additional coordinates or degrees of freedom to describe the spatial extension of a lateral chain has not been investigated and might not be dramatic.

Investigations concerning the modeling of polypeptide chains have appeared recently in the literature. The problem of the reliability of lattice models for protein modeling has been addressed by Godzik et al. (1993), who set resolution limits for a wide number of lattices. Furthermore, the statistical properties of proteins when represented as C_α have been investigated by Oldfield and Hubbard (1994), who studied in detail the correlation between pseudo-bend and pseudo-dihedral angles involving four consecutive C_α atoms, and by De Witte and Shakhnovich (1994), who correlate the pseudo-dihedral angle with the types of the two central residues and use the corresponding distributions to derive torsional potentials in the quasichemical approximation.

The distributions of $C_{\alpha i}$–$C_{\alpha i+1}$ distances (subscript i indicates the ith residue), $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$ bend angles, and $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ torsion angles was also used by Rackovski (1990) to classify protein structures at a short length scale, although it was suggested that similar concepts might be employed at a larger length scale.

In this work we focus on simplified models of polypeptide chains for global conformational search. We investigate the expected and observed behaviors of polypeptide chains when modeled as C_α chains and C_α chains with attached C_β's. We also propose to model side chains via rotational ellipsoids, and we investigate two different forms of interaction potential to model side chain–side chain van der Waals energy. Although other energy terms may be implemented in a way similar to those adopted so far, a non-spherical symmetry model for side-chain steric hindrance requires some discussion.

As is well known, polypeptide chains are chains of covalently linked amino acids. Each amino acid is characterized by its lateral chain, whereas, with few exceptions, the backbone of all amino acids spans more or less the same conformational space. The most detailed description within the framework of classical force fields is the so-called all-atom model where each atom of the molecule is described by a set of coordinates and atomic properties such as charge and Lennard-Jones parameters. A table of covalent linkages is also given, and a force field that usually includes bond, bend angle, torsion angle, non-bond (van der Waals and electrostatic), and hydrogen-bond energy terms is used to assign to any conformation its energy. It is obvious that, even much beneath typical protein size, such a detailed description is not suited for any systematic conformational search. Simplified models have therefore been purposely designed to perform such searches. There is a tradeoff between the details included in the model and the computational effort needed to evaluate the energy of a conformation. Molecular details also increase, in a less severe way, the time needed to generate a conformation. There is also a tradeoff between the width of the phase space one considers for each degree of freedom and the number of conformations one can explore. All-atom models and lattice models may be viewed as the diverging ends of a hierarchy whose steps progressively evolve in a coarser-grained or a more rigid picture of the molecule. For instance, neglecting all of the hydrogen atoms of a molecule, and accordingly readjusting the force field, is a coarse-graining step whose aim is to reduce the time needed to generate a conformation and to evaluate its energy. Similarly, setting bond lengths to constant values or restraining torsion angles to a few fixed values is a stiffening step whose aim is to reduce the number of generated structures.

The first question we raise is, what is the behavior of polypeptide chains when the crudest model is used, i.e., when only C_α's are considered? We will consider therefore pseudo-bonds connecting C_α's of adjacent residues. The database distributions are compared with those obtained by global conformational search employing a stretch of two or three alanines. The comparison highlights local versus nonlocal features of the distributions.

Once this behavior has been described, we pose the problem of including asymmetry due to side chains. Attaching the C_β carbons to the C_α chain appears a good choice because the orientation of the C_β–C_α bond may be representative of the direction of the side chain with respect to the main chain. Indeed, Skolnick and Kolinski (1990b) have already succesfully used such a representation of a polypeptide chain on a lattice.

We further investigate whether additional degrees of freedom, like a pseudo-bond bend angle or torsion angle, are needed for the description of the behavior of the C_β carbons or whether their position may be taken as fixed with respect to the C_α chain. Rey and Skolnick (1992) have shown that the position of the C_β carbon may be reconstructed to a high degree of accuracy by the arrangement of the attached C_α and the two flanking C_α carbons. Compared to their approach our description aims at maintaining the C_α chain pseudo-dihedral angles as the only degrees of freedom.

Lattice nodes or C_α positions have been used to simulate residue properties via spherical potentials. We ask whether it is possible to include in the model the description of the elongate shapes of side chains via rotational ellipsoids, which appear to be the most simple geometrical shapes that preserve orientation in space.

METHODS

Software and database

To search systematically the conformational space of small stretches of a polypeptide chain we have used an efficient program written in C and implemented on machines running under UNIX, DOS, and OS/2 operating systems. The program, which is the reserved property of Italfarmaco S.p.A., was developed by one of us (SC) (Cattarinussi, 1994).

The program performs a tree-based conformational search over a library of fragments created by the program itself. The starting point of the whole process is an input file that defines the topology and the energy parameters of the primitive fragments.

A primitive fragment is defined by a group of atoms whose relative positions are conserved in all of the molecule conformations and whose space positions depend on the same set of torsional angles. Obviously, such primitive fragments exist only under the assumption of constant bond lengths and bond angles.

According to this definition, the following four primitive fragments should be defined to generate the conformations of a very simple chain such as: $C^aH^a_3$–$C^bH^b_2$-$C^cH^c_3$: $f_1 \rightarrow (H^{1a}, C^a)$, $f_2 \rightarrow (H^{2a}, H^{3a}, C^b)$, $f_3 \rightarrow (H^{1b}, H^{2b}, C^c)$, and $f_4 \rightarrow (H^{1c}, H^{2c}, H^{3c})$. With this choice, it is assumed that the chain is grown from atom C^a toward atom C^c.

For instance, when adding fragment f_4 to an already generated portion of the molecule, the only needed additional information concerns the position of the atoms of that fragment expressed in a suitably chosen local reference frame. The space coordinates for the added atoms are then calculated by simply applying a coordinate transformation. Obviously, a set of local coordinates should be available for every allowed value of the torsion angle, φ, around the C^b–C^c bond. The collection of these coordinate sets forms what we call a "catalog" for the primitive fragment f_4. We note that the catalog for fragment f_1 should contain only one set of coordinates because the coordinates of atoms (H^a, C^a) do not depend on any torsion angle.

The primitive fragment library, that is, the ensemble of all catalogs for primitive fragments, is generated first and can then be used to either directly generate the various conformations of the molecule of interest or to construct catalogs for larger fragments, which corresponds to creating a higher level library. For instance, a higher level catalog could be formed by the conformations of compound fragment $F \rightarrow (f_1, f_2)$. The number of library levels is only limited by the available memory. Libraries of various levels can be used at any time provided they have already been generated.

There are three advantages in generating and using multilevel libraries. First, they simplify the writing of the program input file. Second, the speed of the program is increased, as many coordinates and energy terms for a chain are precalculated. Finally, they allow for an "energy-based" selection process of the fragment conformations, which are inserted in the various catalogs. Such a selection is possible because a weighting function (energy) is evaluated for each nonoverlapping conformation. The selection restricts in a combinatorial fashion the number of generated structures.

The very large changes that can occur between successive conformations precluded us from using Verlet lists (Verlet, 1967) that would have been updated at each step. We used, instead, a procedure called "range search" (Sedgwick, 1990) that is based on a binary tree organization of the atom coordinates.

The program can handle more than one molecule. This feature allows for the analysis of the interactions between two or more linear molecules and, perhaps most important, allows for the construction of branched molecules (each branch is seen as a single molecule whose overall position and orientation depend on the conformation of the "parent" molecule).

Among other features of the program that will be described in a forthcoming paper, we mention the ability to search the conformational space of a molecule in a nonhomogeneous environment (the presence of a cell membrane or a large protein domain) and the possibility of imposing distance constraints.

For the purposes of the present work, as far as the ouput is concerned, the program works as many other molecular mechanics packages.

We have employed in all the calculations the AMBER forcefield (Weiner et al., 1984, 1986) with the bond fixed lengths being taken from the AMBER residue library as supplied by the Biosym software (Biosym Technologies, San Diego, CA). In all calculations 1__4 interactions were scaled by a factor 0.5, and a distance-dependent dielectric constant of $4r$ (r expressed in Å, 1 Å = 0.1 nm) was used.

To monitor the statistical behavior of simplified models of polypeptide chains we have used a restricted structural database of 75 high-resolution (<2.5 Å) protein structures (PDB id. code: 1bov A, 1cob A, 1csc, 1cse E, 1cse I, 1f3g, 1fkf, 1gd1, 1gst A, 1hoe, 1ifc, 1ipd, 1lfi, 1mbc, 1msb A, 1nsb A, 1ova A, 1paz, 1phh, 1rbp, 1rnh, 1tpk A, 1ubq, 1 utg, 1ycc, 256b A, 2aza A, 2ca2, 2cdv, 2cna, 2cpp, 2cyp, 2er7 E, 2fb4 H, 2fcr, 2gbp, 2hip A, 2liv, 2ovo, 2rhe, 2sic I, 2sn3, 2trx A, 2ts1, 2tsc A, 3b5c, 3bcl, 3blm, 3chy, 3cox, 3ebx, 3grs, 3lzm, 3sdp A, 4bp2, 4cla, 4cpv, 4dfr A, 4enl, 4fgf, 4gcr, 4icb, 4ptp, 5cpa, 5p21, 5pti, 5rxn, 6ldh, 6tmn E, 7aat A, 7rsa, 8acn, 9pap, 9rnt, 9wga). The list was obtained from the Brookhaven Protein Data Bank user-group directory (internet address: ftp.pdb.bnl.gov) and first published in Williams et al. (1994). The 75 protein chains have little sequence or structural similarity (sequence identity of <30% and structure sequence alignment program analogy score of <80) and are representative of several protein families (Orengo et al., 1993), so that they may be considered an unbiased sample of the whole databank. The five highest resolution structures that did not contain disulfide bridges (PDB id. code: 1cse I, 1 utg, 1ycc, 5p21, 5rxn) were further chosen to test the ellipsoid-ellipsoid potential. Hydrogens were added to the structures within the program Insight II (Biosym Techonologies, San Diego, CA), and 300 minimization steps were performed to remove the few high-energy spots that could result in artifacts in the analysis.

Van der Waals interactions among all pairs of residue side-chain atoms (i.e., excluding the backbone atoms) were calculated with the program Discover (Biosym Technologies, San Diego, CA).

Tests were also performed on thyroid trancription factor 1 homeodomain (TTF-1 HD). The structure used had previously been obtained by homology modeling and has recently been confirmed by NMR (Fogolari et al., 1993; Viglino et al., 1993; Esposito et al., unpublished results). We choose this protein as a test case because a clear picture of relevant interactions within the homeodomain has emerged in recent years and we have had experience with the structural details of TTF-1 HD. Moreover, several experimental dynamic determinations ranging from the atomic to the molecular level are being accessed via NMR, so that the same protein domain might be used as a model for further developments of the methods proposed in this communication.

C_α chains

When replacing a real polypeptide chain with a pseudo-chain consisting of linked atoms, the properties of this pseudo-polymer should be investigated to efficiently design the lattice to be used to model the chain or to properly bias the conformational search. For instance, the observation that the pseudo-bond angle among three consecutive C_α's lies in the range of 85° to 145° enabled Delisi and co-workers (Brower et al., 1993) to discard 16 of 24 possible moves for each chain step on a lattice. The same observation was used in a more general way by Kolinski and Skolnick (1994a) to restrict the number of lattice moves for a polypeptide chain.

A chain may be characterized in the most basic way by the distributions of the pseudo-bond distances, bend angles, and torsion angles (Fig. 1). The latter parameter distribution may introduce elements of asymmetry that are usually absent from lattice models of polymers and distance-based force fields; in other words, a configuration has exactly the same energy as its mirror image (note that a similar problem has been found for experimental structures deposited in the database; Pastore et al., 1991).

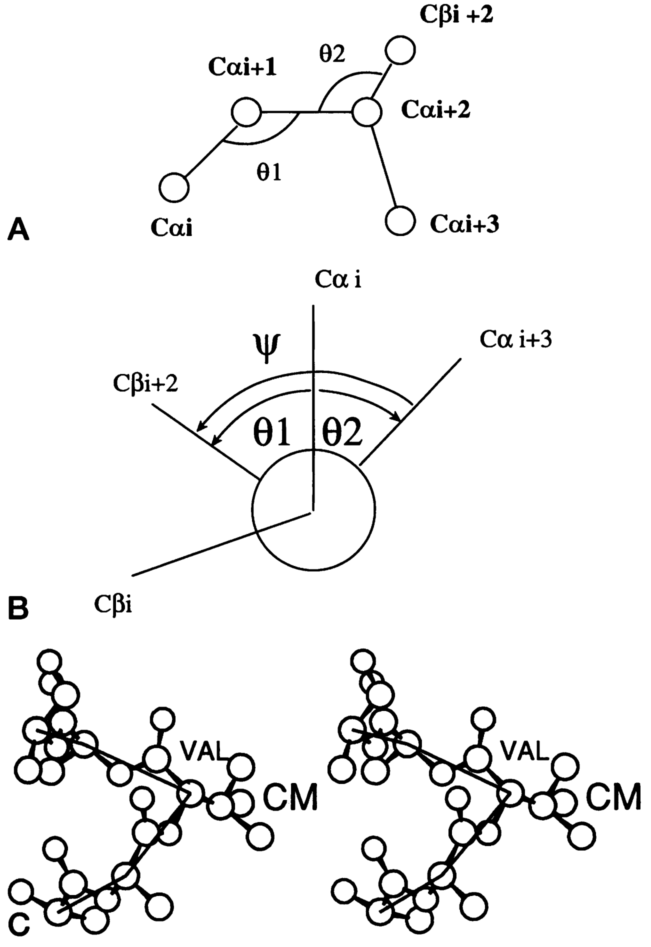
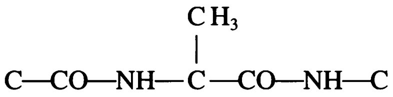


FIGURE 1 (*a*) The pseudo-bond bend angles defined by $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$ (θ_1) and $C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\beta i+2}$ (θ_2) are shown. (*b*) The pseudo-bond torsion angles defined by $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ (θ_2) and $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\beta i+2}$ (θ_1). The view is along $C_{\alpha i+1}$–$C_{\alpha i+2}$ pseudo-bond. The clockwise arrow defines positive angles and the anticlockwise arrows indicate negative angles. The angle ψ obtained by the subtraction ($\theta_1 - \theta_2$) is defined as the $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\beta i+2}$ pseudo-bond torsion angle offset with respect to the $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ torsion angle. (*c*) Stereo view of the center of steric mass (CM) for a valine residue in a stretch of alanines. The C_α pseudo-chain is shown together with all the heavy atoms.

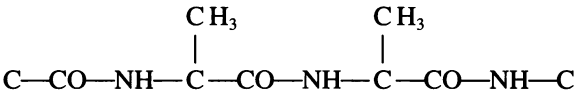
As the distance between two consecutive C_α's is highly restricted near its equilibrium value, we may assume this to be constant, so that the C_α–C_α pseudo-bond, for the usual *trans*-peptide bond geometry, has a definite length (~3.78 Å). Residues found in the *cis* geometry (mainly, but not exclusively, prolines) were excluded from the calculations for consistency.

To analyze database distributions a reference distribution was calculated by global conformational space exploration for two small model compounds. In particular, we have taken a "peptide" stretch including one or two alanines, which, with the neglect of glycines, were considered as minimal side-chain residues, and we have rotated the relevant intervening ϕ and ψ angles in steps of 10°. All of the atoms whose coordinates could depend on the intervening main-chain dihedral angles were included. All of the bond lengths, bend angles, and remaining torsion angles were kept constant at the values found in the AMBER residue library as supported by the Biosym software package Discover. For the sake of clarity the stretch of molecule that was considered for the pseudo-bond bend angle distribution was



$$\mathrm{C{-}CO{-}NH{-}\overset{CH_3}{\overset{|}{C}}{-}CO{-}NH{-}C}$$

and 36 × 36 conformations were generated, whereas for the pseudo-bond torsion angle distribution the stretch was



$$\mathrm{C{-}CO{-}NH{-}\overset{CH_3}{\overset{|}{C}}{-}CO{-}NH{-}\overset{CH_3}{\overset{|}{C}}{-}CO{-}NH{-}C}$$

and 36 × 36 × 36 × 36 conformations were generated.

The pseudo-bond bend angle and the energy of each conformation were then evaluated (the energy did not include constant energy terms). After this each conformation was given a Boltzmann weight ($kT = 0.6$ kcal/mol, 1 cal = 4.184 J). The pseudo-bond bend angle range was divided into one-degree-wide bins, and for each bin the sum of the corresponding weights was evaluated and thus, after normalization, the distribution was obtained.

The other distribution crucial to the behavior of the chain is that of the torsion angle among four consecutive C_α's. The same kind of analysis depicted above for the pseudo-bond bend angle has been performed for this variable. However, we retained only the 15,503 lowest energy generated structures to match the number of available torsion angles in the database.

As glycines and prolines exhibit distinct conformational propensities, distributions of bend and torsion angles were generated for all the stretches that had a proline or a glycine at the first, second, third, or fourth position. Although the distributions are different from the global ones, this observation does not allow a clear-cut partition of the range of possible values. For this reason we did not discriminate between glycines and prolines.

C_α, C_β chains

The C_α–C_β bond orientation is obviously related to the orientation of the side chains with respect to the main chain of the protein. Aiming at maintaining a minimal description of the protein we may ask whether C_β's may be introduced into the C_α chain, so that the torsion angles around the C_α–C_α pseudo-bond still remain the only degrees of freedom of the chain, but, nevertheless, side-chain parameters are introduced through the C_β atom rather than through C_α. This possibility would allow for a more realistic description of each amino acid with a very limited increase in computational load. As the C_α–C_β distance is almost constant, we may easily define the position of the C_β with respect to the main chain in terms of the bend angle defined by $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$ and the torsion angle defined by $C_{\alpha i-1}$–$C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$. Because the latter may depend on the local conformation of the chain, the offset of the same angle with respect to the $C_{\alpha i-1}$–$C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$ dihedral angle (i.e., the dihedral angle obtained by the subtraction $\theta(C_{\alpha i-1}–C_{\alpha i}–C_{\alpha i+1}–C_{\beta i+1}) - \theta(C_{\alpha i-1}–C_{\alpha i}–C_{\alpha i+1}–C_{\alpha i+2})$) appears to be a more useful quantity to monitor (Fig. 1). The theoretical distribution was computed in the same fashion as described above.

One may also wonder to what extent the C_β position is representative of the real side chain. To address the issues of the center and distribution of the global steric hindrance of a group of atoms, we will use some concepts used in classical mechanics to describe mass distributions. To accomplish the parallel with classical mechanics we assign to each atom i a "steric mass" (expressed in arbitrary units) that is equal to its steric hindrance as calculated using its van der Waals radius, a_1, i.e., $m_i = 4\pi a_i^3/3$. It must be clear that no real mass is involved in this definition, and the word "mass" is used for consistency with the formalism developed for the description of distributions of real masses. The word "weight" might be equally appropriate for the purpose of the computation of the center of volume; however, the idea of mass leads easily to that of density, which will be used later. Note that a similarly ad hoc defined center of mass, rather than the real center of mass, has been used previously by Kolinski and Skolnick (1994a) to locate centers for side-chain interactions.

Given these definitions we can compute for every side chain in a defined conformation the center of steric mass: $\vec{R} = \sum_i m_i\vec{r}_i/M$, where $\vec{r}_i$ is the position of the ith atom and $M = \sum_i m_i$ is the global steric mass of the residue (Fig. 1). The angle between the vectors joining C_β to C_α and the center of steric mass to C_α gives an indication of how close the two directions are.

We have computed the distribution of the angle, making use of our restricted protein structures set. Because hydrogen or deuterium atoms are not included in most of the structures these atoms were not included in the calculation for consistency. In addition, incomplete side chains have not been considered. We have also computed the distribution of the pseudo-bond bend angle and torsion angle offset for the center of steric mass, in the same fashion as described for the C_β carbon.

The choice of a steric mass might seem arbitrary, but we believe it is suited to the purpose of modeling because it is precisely the steric hindrance that often makes it possible to select regions of conformational space. We have used the following average values for the steric mass of each atom (in arbitrary units):

C	steric mass = 28.73	(van der Waals radius = 1.9 Å)
N	steric mass = 14.14	(van der Waals radius = 1.5 Å)
O	steric mass = 11.49	(van der Waals radius = 1.4 Å)
S	steric mass = 26.52	(van der Waals radius = 1.85 Å)

C_α and ellipsoids

The previous discussion directly introduces the issue of a more realistic representation of the side chains or (at an increasing level of approximation) parts of the side chains.

The idea of modeling small molecules with ellipsoids is not new and has proved useful in the field of molecular simulation of liquids (Allen and Tildesley, 1987). Often, when a small molecule is depicted with a space-filling model, it has a curved and elongate shape. The most simple shape that can be fit to such a picture is a rotational ellipsoid.

To build a rotational ellipsoid that can suitably represent a set of atoms we impose the requirement that both the original set of atoms and the ellipsoid representing them have roughly the same steric hindrance, the same position, and the same radial distribution around any axis in space, in a sense that is defined below.

With this aim, we defined in the previous section the steric mass of an atom as the volume it occupies in space, according to its van der Waals radius. When such a definition of mass is given, other properties may be easily derived, besides the center of steric mass. For the purpose of deriving the ellipsoid parameters, we assume the steric mass of each atom to be uniformly distributed over the volume. It must be clear that this assumption does not have or hint at any physical meaning, but it just serves the purpose of deriving the dimensions of the ellipsoid. Dealing with a fit involving surfaces, instead of volumes, would probably not require any similar assumption but would certainly lead to much more complex equations, with probably little advantage in the present context.

The asymmetrical distribution of point masses around their center of mass may be suitably represented via the inertia tensor, whose definition is the following (Goldstein, 1965):

$$\mathbf{I} = \sum_{\mathrm{i}} m_{\mathrm{i}}[(\vec{r}_{\mathrm{i}} \cdot \mathbf{1} \cdot \vec{r}_{\mathrm{i}}) - (\vec{r}_{\mathrm{i}}\vec{r}_{\mathrm{i}})]$$

where the dot indicates a scalar product, **1** is the unit matrix, and the dyadic notation $(\vec{a}\vec{b})_{\mathrm{ij}}$ has the following meaning: $(\vec{a}\vec{b})_{\mathrm{ij}} = a_{\mathrm{i}}b_{\mathrm{j}}$. In the present context we will use the word "tensor" as synonymous with "matrix," although in certain respects the two words have different meanings. We keep the wording "inertia tensor" to be consistent with classical mechanics textbooks. Besides the dynamical properties that render the inertia tensor so important in the field of classical mechanics, we note that it also possesses a powerful geometrical meaning, as it condenses all the information about radial distribution of mass around any axis. Together with the coordinates of the center of mass, it is therefore suited to representing the position and the dispersion of a distribution of mass.

The inertia tensor can be diagonalized to obtain the principal axis. The three eigenvalues one obtains, even if they may incidentally be coincident, are the inertia momenta with respect to a set of three orthogonal axes whose directions are given by the correponding eigenvectors.

Once the center of steric mass and the steric mass inertia tensor of the original set of atoms have been obtained, we choose, as a representation of the original set of atoms, the rotational ellipsoid that possesses the same steric mass, the same center of steric mass, and the same (rotationally averaged) steric mass inertia tensor as the original set of atoms.

We assume the steric mass of the ellipsoid, defined as before as being equal to the volume the ellipsoid occupies in space, to be uniformly distributed in space. The center of steric mass is obviously the center of the ellipsoid itself, whereas the steric mass momenta of inertia along the axis of rotation and any orthogonal axis may be obtained by direct integration of the radial distribution of steric mass.

It turns out that for a rotational ellipsoid whose mass M is uniformly distributed within an ellipsoidal surface with the major axis of length b, along the rotation axis, and the minor axis of length a, we have the following properties:

i) the equation representing its surface is

$$(\vec{r} - \vec{R}) \cdot \gamma^{-1} \cdot (\vec{r} - \vec{R}) = 1,$$

where $\vec{R}$ is the position of the ellipsoid center of mass, and the matrix γ^{-1} may be expressed in terms of the length of the axes and the unit vector $\vec{v}$ along the rotation axis:

$$\gamma^{-1} = \frac{1}{b^2}\vec{v}\vec{v} + \frac{1}{a^2}(\mathbf{1} - \vec{v}\vec{v}).$$

When $\vec{R}$ coincides with the origin and $\vec{v}$ coincides with the z axis the equation for the surface reduces to the familiar form

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} = 1;$$

ii) its mass density is

$$\rho = \frac{M}{(4\pi/3)ba^2};$$

iii) the momenta of inertia with respect to the axis of rotation (I_{p}) and any orthogonal axis through the center of mass (I_{t}) are

$$I_{\mathrm{p}} = \rho \int_{-\mathrm{b}}^{\mathrm{b}} \mathrm{d}z \int_{0}^{\mathrm{a}\sqrt{(1-z^2/b^2)}} r^2 2\pi r\,\mathrm{d}r = \frac{2}{5}Ma^2$$

$$\text{with} \quad r = \sqrt{x^2 + y^2}$$

and

$$I_{\mathrm{t}} = \rho \int_{-\mathrm{b}}^{\mathrm{b}} \mathrm{d}z \int_{0}^{2\pi} \mathrm{d}\vartheta \int_{0}^{\mathrm{a}\sqrt{(1-z^2/b^2)}} r(r^2 \sin\vartheta^2 + z^2)\,\mathrm{d}r$$

$$= M\left(\frac{1}{5}a^2 + \frac{1}{5}b^2\right)$$

$$\text{with} \quad r = \sqrt{x^2 + y^2 + z^2},$$

respectively.

Because the inertia tensor of the original distribution of steric mass, contrary to that of a rotational ellipsoid, possesses in general three different eigenvalues, we first have to average the two larger eigenvalues to obtain an average momentum of inertia that is transverse with respect to the remaining axis, i.e., the axis about which rotational average has been performed. Then we can simply equate i) the steric mass, ii) the steric mass parallel and transverse momenta of inertia, and iii) the direction of the rotation axis for the original set of atoms and the ellipsoid, so that the rotational ellipsoid is equal, under these three respects, to the original set of atoms.

We may clarify the procedure and its implications by considering two atoms with a van der Waals radius *r*, placed at a distance of 3/2 *r* from each other (Fig. 2). The global steric mass is $M = 2(4\pi/3)r^3$. The center of steric mass of the system is placed at the midpoint of the vector joining the two atomic centers of steric mass. The steric mass inertia tensor may be easily calculated by taking a Cartesian coordinate reference system that has the *z* axis aligned with $\vec{v}$ and an origin coincident with the center of steric mass. Each entry of the steric mass inertia tensor is composed of two terms, one describing the atoms as steric point masses, and the other describing the steric mass distribution of each atom. In this simple case the inertia tensor is the following:

$$\mathbf{I} = \begin{pmatrix} M\left[\left(\frac{3r}{4}\right)^2 + \frac{2}{5}r^2\right] & 0 & 0 \\ 0 & M\left[\left(\frac{3r}{4}\right)^2 + \frac{2}{5}r^2\right] & 0 \\ 0 & 0 & M\frac{2}{5}r^2 \end{pmatrix}.$$

The three diagonal elements of the tensor are obviously the eigenvalues, two of which are coincident. If the Cartesian coordinate system were chosen differently, then we would also have nonzero off-diagonal terms, and diagonalization of the matrix would be required. Note that the smaller eigenvalue is associated with the direction of elongation. If the two larger eigenvalues were not coincident we would have simply taken their average. Now we consider which ellipsoid would better fit our two-atom system, according to the previous requirements. If *a* and *b* are the minor and major ellipsoid axes, respectively, and the major axis is along the rotation axis, then i) the steric mass of the ellipsoid is *M* and the center of the ellipsoid coincides with the center of steric mass of the two atoms; ii) from

$$I_{\mathrm{p}}\left(= M\frac{2}{5}a^2\right) = M\frac{2}{5}r^2$$

we obtain $a = r$; iii) from

$$I_{\mathrm{t}}\left(= M\left(\frac{1}{5}b^2 + \frac{1}{5}a^2\right)\right) = M\left[\left(\frac{3r}{4}\right)^2 + \frac{2}{5}r^2\right]$$

we obtain $b = 1.95r$. One can appreciate the fair fit of the ellipsoid to the two atoms (Fig. 2). The fit is going to be less satisfactory with an increasing degree of asymmetry, but this representation should work fairly well with the short (with respect to bond lengths and atomic radii) chains of amino acids.

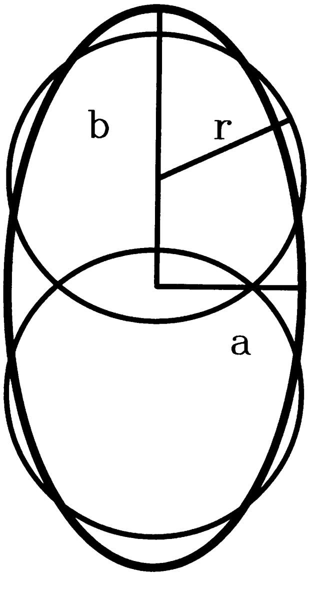


FIGURE 2 A two-atom system represented as a rotational ellipsoid.

We have tested whether such a procedure gives a reasonable representation of amino acids that are quite far from possessing any rotational symmetry.

The issue has been preliminarily addressed by scanning the set of 75 high-resolution protein structures and considering the statistical distributions that are obtained for

i) the dimensions of the ellipsoids representing each amino acid;

ii) the angle between the rotation axis of the ellipsoid and the C_α center of steric mass direction, or the C_α–C_β direction, which should reasonably represent the direction of the side chain in space.

The latter parameter, however, has a different relevance, depending on the asymmetry of the ellipsoid, which ultimately reflects the directionality of the side chain.

All chains for which atoms were missing were excluded from the count for consistency, and hydrogen or deuterium atoms were not taken into account, as most of the database includes only heavy atoms. More accurate dimensions should therefore include some additional length corrections for the hydrogen atoms.

A good test for the usefulness of ellipsoids for molecular simulation is to write an interaction potential that is able to reproduce energy terms obtained with all-atom models. Developing a force field for ellipsoid-ellipsoid interactions may be a complex task because of the variables (axis lengths and orientations with respect to the vector joining the centers of mass for both ellipsoids) that determine the energy of the interaction, whichever interaction model is assumed. The modeling of ellipsoid-ellipsoid interactions is definitely beyond the scope of the present work. However, based on simple geometric and physical considerations, we have written two simple forms for the interaction potential that give reasonable results for simple situations. We focus on the van der Waals potential, as this is often used to reject possible conformations in systematic conformational searches and because this presents the major differences with respect to spherical models of side chains. More detailed force fields accounting for backbone and side-chain interactions have appeared recently in the literature (Gerber, 1992; Kolinski and Skolnick, 1994a).

We have modeled the van der Waals interaction of two ellipsoids with a Lennard-Jones potential retaining the underlying physical interaction. The spherical 6–12 potential between two atoms i and j may be easily generalized to an ellipsoidal potential by substituting r_{ij}^2/r_0^2 with $\vec{r}_{ij}\gamma^{-1}\vec{r}_{ij}$. According to this picture we have assigned the following functional form to the interaction potential, which uses an average ellipsoid given by a simple function of the two interacting ellipsoids:

$$V_{ij} = \epsilon_{ij}((\vec{r}_{ij}\gamma_{ij}^{-1}\vec{r}_{ij})^{-6} - (\vec{r}_{ij}\gamma_{ij}^{-1}\vec{r}_{ij})^{-3}).$$

The choice of ϵ_{ij} and γ_{ij}^{-1} defines the Lennard-Jones potential properties. If the tensor γ_{ij}^{-1} or ϵ_{ij} is made to depend on the orientation of $\vec{r}_{ij}$, then isopotential contours may possess any shape. In the simplest choice, which follows the original idea of Berne and Pechukas (1972), the average matrix, which defines isopotential energy surfaces, is obtained as a simple function of the corresponding matrices of the two interacting ellipsoids, i.e., $\gamma_{ij}^{-1} = k^2(\gamma_i + \gamma_j)^{-1}$. The expression stems from the observation that, when k is set to 1, $\exp(-\vec{r}_{ij}\gamma_{ij}^{-1}\vec{r}_{ij})$ is proportional to the overlap of two gaussian ellipsoidal mass densities defined by γ_i^{-1} and γ_j^{-1} (Berne and Pechukas, 1972). For this reason we will refer to this force field as the "overlap model," although it is not strictly coincident with the model of Berne and Pechukas (1972). This choice is very simple in form, and the overlap allowance is set by the value of k, which divides the linear dimensions of the repulsive core of the ellipsoid, maintaining its shape. For two equal interacting ellipsoids a choice of $k = 0.707$ would set the dimensions of the ellipsoid two times larger than the original ones. This would be the usual choice for two interacting spheres. This potential choice is bound to break down when ellipsoids of largely different sizes are made to interact.

Optimal values for ϵ_{ij} and k were determined from the set of five high-resolution structures. Ellipsoids representing the side chains were obtained from the actual positions of the side-chain atoms. A value of 0.81 for k was determined by the analysis of the dependence of both total energy and number of clashes (positive energy contacts) on k in the set of five high-resolution structures after energy minimization. When one or both of the interacting residues were aromatic, this value was further multiplied by a factor of 1.09 or 1.19, respectively. With this choice no clash is found.

ϵ_{ij} has been chosen to be proportional to the volume of both the interacting ellipsoids, and its absolute value was determined by a best fit to all residue pair van der Waals energies above 0.001 kcal/mol. For two interacting alanines ϵ_{ij} is 1.21 kcal/mol, not too far from the value of 0.72 kcal/mol found in united atom force fields like, for instance, GROMOS (Van Gunsteren and Berendsen, 1987). A list of the energy parameters ϵ_{ij} for all pairs of standard amino acids is given in Table 1.

Other functional forms have been proposed that can much more faithfully reproduce the interaction energy of clusters of atoms (Gay and Berne, 1981) but require, on the other hand, quite a larger computational effort. A problem in attaining a faithful representation of side chain–side chain van der Waals interaction is that small adjustments of the orientations or dimensions of the ellipsoids may turn high-energy clashes into favorable interactions. One should

TABLE 1 **Energy parameter ϵ_{ij} (kcal/mol) for standard amino acid pairs for the "overlap model"**

	ALA	CYS	ASP	GLU	PHE	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	1.21	1.72	1.88	2.39	3.63	2.74	2.72	2.90	2.74	2.94	1.95	2.09	2.47	3.64	1.26	1.79	2.12	4.74	3.74
CYS	1.72	2.45	2.67	3.41	5.16	3.90	3.87	4.12	3.90	4.18	2.77	2.98	3.52	5.18	1.79	2.54	3.01	6.74	5.32
ASP	1.88	2.67	2.91	3.71	5.63	4.25	4.23	4.50	4.26	4.56	3.02	3.25	3.84	5.65	1.95	2.77	3.28	7.35	5.80
GLU	2.39	3.41	3.71	4.74	7.18	5.42	5.39	5.74	5.43	5.82	3.86	4.15	4.89	7.21	2.49	3.54	4.19	9.38	7.40
PHE	3.63	5.16	5.63	7.18	10.88	8.21	8.17	8.69	8.23	8.81	5.84	6.28	7.41	10.92	3.78	5.36	6.35	14.21	11.21
HIS	2.74	3.90	4.25	5.42	8.21	6.20	6.16	6.56	6.21	6.65	4.41	4.74	5.60	8.25	2.85	4.05	4.79	10.73	8.46
ILE	2.72	3.87	4.23	5.39	8.17	6.16	6.13	6.52	6.18	6.61	4.39	4.72	5.57	8.20	2.84	4.02	4.76	10.67	8.41
LYS	2.90	4.12	4.50	5.74	8.69	6.56	6.52	6.94	6.58	7.04	4.67	5.02	5.92	8.73	3.02	4.28	5.07	11.35	8.96
LEU	2.74	3.90	4.26	5.43	8.23	6.21	6.18	6.58	6.23	6.67	4.42	4.75	5.61	8.27	2.86	4.05	4.80	10.75	8.48
MET	2.94	4.18	4.56	5.82	8.81	6.65	6.61	7.04	6.67	7.14	4.73	5.09	6.00	8.85	3.06	4.34	5.14	11.51	9.08
ASN	1.95	2.77	3.02	3.86	5.84	4.41	4.39	4.67	4.42	4.73	3.14	3.37	3.98	5.87	2.03	2.88	3.41	7.63	6.02
PRO	2.09	2.98	3.25	4.15	6.28	4.74	4.72	5.02	4.75	5.09	3.37	3.63	4.28	6.31	2.18	3.09	3.66	8.20	6.47
GLN	2.47	3.52	3.84	4.89	7.41	5.60	5.57	5.92	5.61	6.00	3.98	4.28	5.05	7.45	2.58	3.65	4.33	9.68	7.64
ARG	3.64	5.18	5.65	7.21	10.92	8.25	8.20	8.73	8.27	8.85	5.87	6.31	7.45	10.97	3.79	5.38	6.37	14.27	11.26
SER	1.26	1.79	1.95	2.49	3.78	2.85	2.84	3.02	2.86	3.06	2.03	2.18	2.58	3.79	1.31	1.86	2.20	4.93	3.89
THR	1.79	2.54	2.77	3.54	5.36	4.05	4.02	4.28	4.05	4.34	2.88	3.09	3.65	5.38	1.86	2.64	3.13	7.00	5.52
VAL	2.12	3.01	3.28	4.19	6.35	4.79	4.76	5.07	4.80	5.14	3.41	3.66	4.33	6.37	2.20	3.13	3.70	8.29	6.54
TRP	4.74	6.74	7.35	9.38	14.21	10.73	10.67	11.35	10.75	11.51	7.63	8.20	9.68	14.27	4.93	7.00	8.29	18.56	14.64
TYR	3.74	5.32	5.80	7.40	11.21	8.46	8.41	8.96	8.48	9.08	6.02	6.47	7.64	11.26	3.89	5.52	6.54	14.64	11.55

choose an interaction potential depending on the purposes of the simulation. It is worth noting that the expression $\vec{r}\gamma^{-1}\vec{r}$ is equal to $(r/d)^2$, where r is the modulus of $\vec{r}$ and d is the distance of the ellipsoid surface from the origin in the direction of vector $\vec{r}$. A potential aiming at maintaining the repulsive cores of the ellipsoids must then take into consideration the ratio $r_{ij}/(d_i + d_j)$ for the Lennard-Jones potential. Indeed, we have chosen this expression for a more accurate energy evaluation. Note, however, that for very asymmetrical ellipsoids, clashes might occur in points that do not lie on the vector joining the two centers of mass. The relationship between $\vec{r}_{ij}\gamma_{ij}^{-1}\vec{r}_{ij}$ and γ_i^{-1} and γ_j^{-1} is derived:

$$\vec{r}_{ij}\gamma_{ij}^{-1}\vec{r}_{ij} = \left(\frac{1}{(\vec{r}_{ij}\gamma_i^{-1}\vec{r}_{ij})^{-1/2}+(\vec{r}_{ij}\gamma_j^{-1}\vec{r}_{ij})^{-1/2}}\right)^2.$$

Because γ_{ij}^{-1} depends on the orientation of $\vec{r}_{ij}$, isopotential contours are not necessarily rotational ellipsoids. Furthermore, here a scaling factor k, which divides the linear dimensions of ellipsoids i and j, sets the tolerance to possible overlaps. A value of 1.12 for k was determined as described for the "overlap model." If one or both of the interacting residues are aromatic residues, k is further multiplied by a factor of 1.09 or 1.20, respectively. ϵ_{ij} should account for the dimensions and mutual orientation of the interacting ellipsoids. ϵ_{ij} has been chosen to be proportional to the cross sections of both the ellipsoids in the direction defined by their centers of steric mass. In this way if two equal ellipsoids have major and minor axes equal to 2 and 1, respectively, the interaction energy of the parallel arrangement is four times that of the sequential arrangement (Fig. 3). ϵ_{ij} was determined as described above, and its value for two interacting alanines is 2.28 kcal/mol, higher than expected and possibly hinting at some inadequacy of the force field. We will refer to this second model as the "repulsive core model." Fig. 3 better illustrates the choices for ellipsoid-ellipsoid van der Waals potential.

A problem that is encountered in determining optimal values for k and ϵ_{ij} is that energy minimization changes both parameters in a consistent way. To avoid clashes and to fit the all-atom energy, both k and ϵ_{ij} must be set slightly larger than they would be for unminimized structures. In other words, minimized structures are consistently more compact and stable than the structures found in the database, and residue-residue favorable contacts are possibly overestimated.

Finally, a test was performed on the same set of five high-resolution proteins using spheres instead of ellipsoids to represent side chains. The radius of the sphere (r) was chosen so as to give the same volume as the corresponding ellipsoid (i.e., with the previous notation, $4\pi r^3/3 = 4\pi l_p l_t^2/3$). The scaling factor $k = 1.07$, analogous to the "overlap model," was determined to avoid clashes. The energy parameters were chosen to be proportional to the volume of the interacting spheres, and their absolute values were chosen to give the best fit to the corresponding AMBER energies. In practice spherical model ϵ_{ij} may be obtained by multiplying by a factor of 1.73 the corresponding values reported in Table 1.

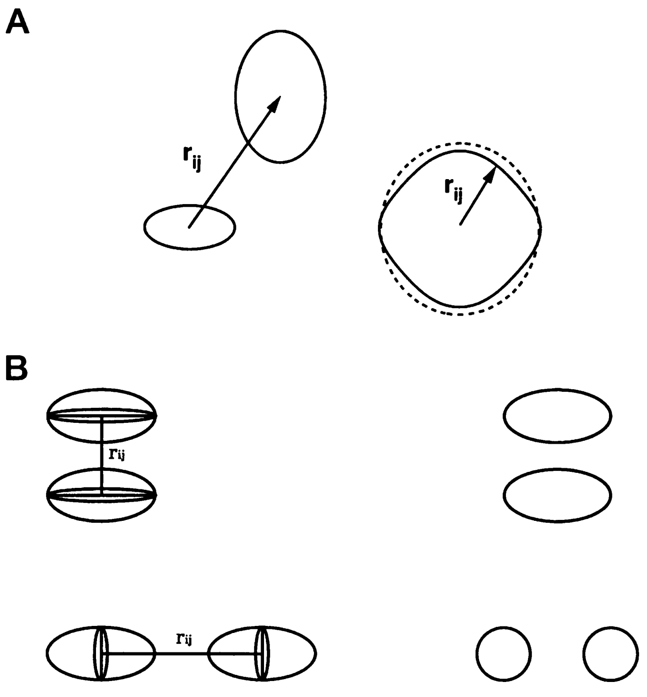


FIGURE 3 Ellipsoid-ellipsoid interaction potential. (*a*) On the left the repulsive cores of two interacting ellipsoids are drawn. On the right the zero isopotential curve is drawn for the "overlap model" (*dashed line*) and for the "repulsive core model" (*solid line*). (*b*) On the left the repulsive cores of two interacting ellipsoids are drawn. On the right the cross section in the direction defined by the two ellipsoid centers is drawn.

RESULTS AND DISCUSSION

C_α chains

The distributions of the pseudo-bond bend angle and torsion angle, when calculated on very short stretches of a polypeptide chain (see Methods), show definite conformational preferences (Figs. 4 and 5).

The bend angle distribution (range 84° to 144°) may be approximately fitted by a gaussian function peaked at a value of 113.51° with a width of 35.16° (twice the standard deviations). The distributions obtained in this way reflect by definition local interactions, so that features that are different in the database distributions are attributable to nonlocal, cooperative, or long-range interactions, like those that may be correlated with helical structures. Indeed, the pseudobond bend angle distribution found in the database (15,663 bend angles) is very similar to the theoretical one (Fig. 4), except for the intense, sharp peak at approximately 90°C, i.e., the characteristic value of α-helical secondary structures. It is worth noting, however, that the distribution may be fitted by two gaussian curves centered at 91.1° and 117.2° and having widths equal to 6.5° and 33.0°, respectively. It is apparent that the second peak has parameters very close to the theoretical single-peak distribution and

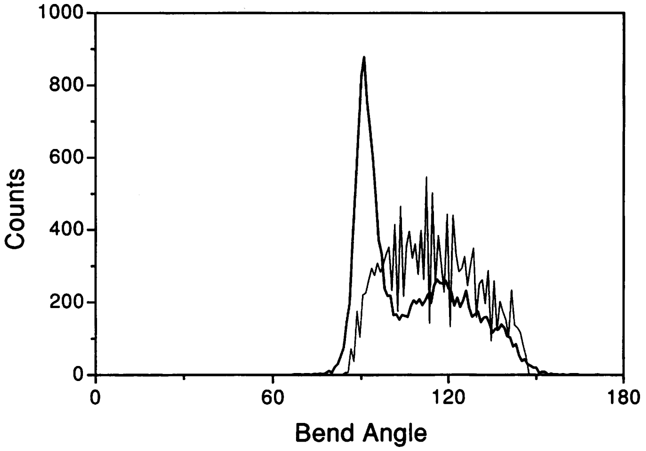


FIGURE 4 $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$ pseudo-bond bend angle distribution. The counts in the database are plotted (*thick solid line*) together with the theoretically computed ones (*thin solid line*). The width of the bins is one degree.

therefore is compatible with locally energy-stabilized conformations, although the contribution of other secondary structure elements stabilized by long-range interactions is expected to fall in the same range of values.

The database range of pseudo-bond bend angle is slightly larger than the calculated one but is limited, reinforcing the idea of a rather rigid pseudo-bond structure of the chain, even though there are six bonds intervening among three consecutive C_α's.

The behavior of the pseudo-bond torsion angle appears much less defined. Although a structure in the calculated distribution is definitely present, the premises for reducing the number of chain moves on a lattice, or any other kind of grid, are greatly lessened. Only a small region of the histogram (around −30°) is poorly populated, so that the reduction in searchable conformational space is of little advantage (Fig. 5).

The presence of an intense and sharp peak around 50° in the distribution computed on the database (15,503 torsion angles) is again related to α-helical secondary structures. When the theoretical distribution is scaled properly to take into account the absence of the peak due to α-helices, which contributes roughly 4000 to 5000 counts, another difference is apparent around 200° that may be related to β-sheets. However, the less strict geometrical requirement of β-sheet versus α-helical conformation makes this peak broad and less readily identifiable. Note that the asymmetrical distribution of dihedral angles may introduce a bias toward, for instance, right-handed α-helices over left-handed ones, a result that cannot be obtained just by distance bias or restraints.

C_α chains with C_β's

The theoretical distribution of the pseudo-bond bend angle defined by $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$ shows a sharp peak, centered at 123° (Fig. 6), and although the distribution is skewed, the standard deviation from the average value (118°) is very small (6°). The corresponding distribution in the database is somewhat broader, but it is still skewed, with a maximum at 125°.

The offset of the torsion angle $C_{\alpha i-1}$–$C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$ with respect to the main-chain torsion angle (defined by $C_{\alpha i-1}$–$C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$) does not show a similar well-defined behavior, but the range of values it can assume is not so large as to prevent any attempt to maintain a rigid geometry for the C_β carbon. The theoretical distribution is bimodal with two peaks roughly at 210° and 250°, which are the typical values for extended and α-helical conformations, respectively, and covers an overall range of approximately 60° (Fig. 7). The same features, but again with a broader peak shape, are exhibited by the distribution found

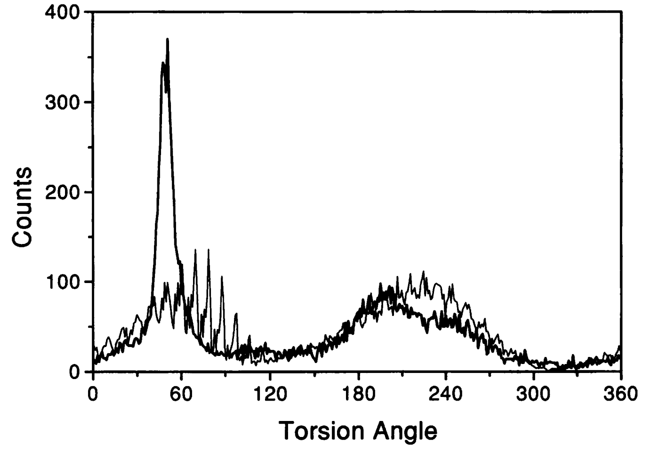


FIGURE 5 $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ pseudo-bond torsion angle distribution. The counts in the database are plotted (*thick solid line*) together with the theoretically computed ones (*thin solid line*). The width of the bins is one degree.

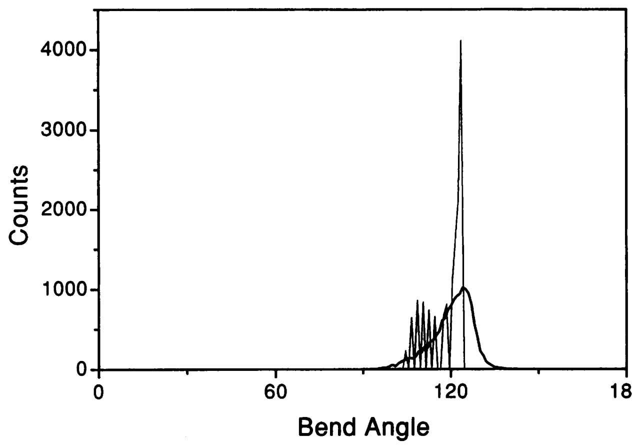


FIGURE 6 $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\beta i+1}$ pseudo-bond bend angle distribution. The counts in the database are plotted (*thick solid line*) together with the theoretically computed ones (*thin solid line*). The width of the bins is one degree.

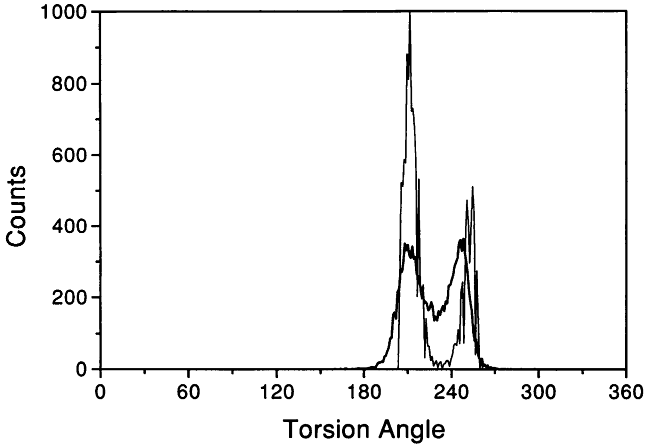


FIGURE 7 Distribution of $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\beta i+2}$ pseudo-bond torsion angle offset with respect to the $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ torsion angle. The counts in the database are plotted (*thick solid line*) together with the theoretically computed ones (*thin solid line*). The width of the bins is one degree.

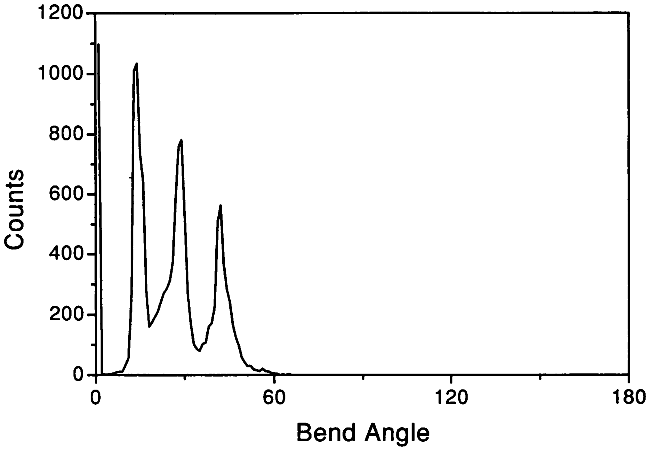


FIGURE 8 Distribution of the angle between the vector joining C_α to the center of steric mass and the bond C_α–C_β. The counts in the database are plotted in one-degree bins.

in the database. Hence we are able to conclude that, assuming a fixed C_α–C_β length, for a simplified description of polypeptide chains, the position of C_β carbons with respect to the corresponding C_α may be described simply by a pseudo-bond bend angle and a torsion angle offset with respect to the main chain constituted by the C_α's. This is the usual way atoms are specified when modeling molecules and assuming rigid bonds and bend angles. When the position of the C_β carbon is specified with respect to the C_α trace by using an average pseudo-bond bend angle of 125° and an average pseudo-bond torsion angle of 245°, the average root mean square deviation from the actual position of the C_β carbon in the set of five high-resolution structures is 0.50 Å. This figure is higher than that obtained by Rey and Skolnick (1992), who, however, used amino acid-dependent parameters to reconstruct the position of the C_β from the local coordinate system defined by three consecutive C_α's. The two approaches, albeit similar, are not equivalent, because the pseudo-bond bend angle is fixed in our approach.

We have extended the analysis further by investigating whether C_β's may adequately represent the side chains of residues. We have examined the angle between the C_α–C_β bond and the vector joining C_α and the center of steric mass (see Methods) to check for the orientation of the bond. The angle shows a trimodal distribution and covers an approximate range of 50°, as can be seen from Fig. 8.

The structure of the distribution is related to the number of rotatable bonds in the side chain, rather than to the specific conformation; i.e., for each side chain the angle between the C_α–C_β bond and the vector joining C_α and the center of steric mass exhibits a limited standard deviation, as reported in Table 2.

The length of the vector joining C_α and the center of steric mass was also evaluated. Its value obviously depends on the examined residue. The average values together with the standard deviations are reported in Table 2. Such lengths should be taken into account when transferring residue properties on the C_β carbon to represent properly not only the orientation but also the position of the side chain. It is worth noting that the standard deviations of the C_α center of steric mass length are small also for long side chains, because folded or kinked conformations, which should exhibit large deviations, are not very common. The same analysis was performed after also including the C_α carbon in the side chain, as this is expected to reduce the angle

TABLE 2 Statistics of the center of steric mass

Residue	Angle (degrees)	Length (Å) C_α not included	 C_α included	Counts
ALA	0.00 (0.0)	1.53 (0.02)	0.76 (0.01)	1355
CYS	23.2 (1.3)	2.03 (0.05)	1.34 (0.03)	329
ASP	28.1 (1.6)	2.24 (0.05)	1.65 (0.04)	945
GLU	27.2 (9.7)	2.83 (0.24)	2.24 (0.19)	872
PHE	42.0 (2.7)	3.41 (0.07)	2.98 (0.06)	610
HIS	39.4 (2.4)	3.01 (0.07)	2.51 (0.06)	370
ILE	16.6 (4.5)	2.34 (0.14)	1.87 (0.11)	815
LYS	29.7 (10.2)	3.29 (0.24)	2.69 (0.20)	1013
LEU	27.6 (2.7)	2.62 (0.07)	2.09 (0.06)	1265
MET	27.6 (11.2)	2.95 (0.27)	2.35 (0.21)	302
ASN	28.7 (1.8)	2.27 (0.06)	1.69 (0.04)	691
PRO	41.2 (0.9)	1.88 (0.03)	1.41 (0.02)	703
GLN	27.5 (9.9)	2.85 (0.25)	2.26 (0.20)	511
ARG	29.4 (13.3)	3.78 (0.30)	3.20 (0.26)	638
SER	12.9 (0.6)	1.71 (0.03)	1.00 (0.02)	1121
THR	15.1 (1.0)	1.94 (0.04)	1.37 (0.03)	931
VAL	13.0 (1.4)	1.97 (0.03)	1.48 (0.02)	1100
TRP	44.3 (8.1)	3.87 (0.19)	3.50 (0.17)	228
TYR	43.7 (2.8)	3.56 (0.08)	3.13 (0.07)	593

Length of the vector joining C_α to the center of steric mass (calculated alternatively including or not the C_α carbon). The angle between the latter vector with the bond C_α–C_β is given for each residue. Averages are given with root mean square deviations in parentheses. Counts for each residue are listed in the last column.

between the longitudinal axis of the rotational ellipsoid, which represents the side chain, and the C_α–C_β (and C_α center of steric mass) direction (see next section).

Finally, as a consistency check we have monitored the distributions of the pseudo-bond bend and torsion angle offset of the residue centers of steric mass with respect to the main chain to see if they match the corresponding distributions obtained for the C_β's. As could be expected, both parameters show a much less well defined behavior than the corresponding quantities involving the C_β's, but the distributions are definitely similar as far as the average values are concerned, with an overall increase in dispersion (Figs. 9 and 10).

C_α and ellipsoids

A prerequisite for a faithful representation of side chains as ellipsoids is to obtain reasonable average values for the parameters of the ellipsoids. It is also necessary that the root mean square deviations of the distributions of these values not be so large as to devalue the significance of the representation.

The distributions of the parameters for each amino acid are not shown here, but they are all approximately bell shaped, with more details for the longer side chains.

A list of the parameters, together with root mean square deviations, is reported in Table 3. It should be noted that the root mean square deviation of the values of the minor axis length are very small, whereas those for the major axis length are somewhat larger. The largest side chains show the highest variability, as expected. The angles between the major axis of the ellipsoid and the vector joining the center of steric mass and the C_α (or the C_β and the C_α) are rather large, as judged by Table 3, so that the latter vectors cannot safely approximate the ellipsoid direction. Nevertheless, when C_α's are included in the calculation of the center of steric mass, the angle between the major axis of the ellipsoid and the C_α–C_β direction (and to a lesser extent also the C_α center of steric mass direction) remains within a very limited range (0–33°).

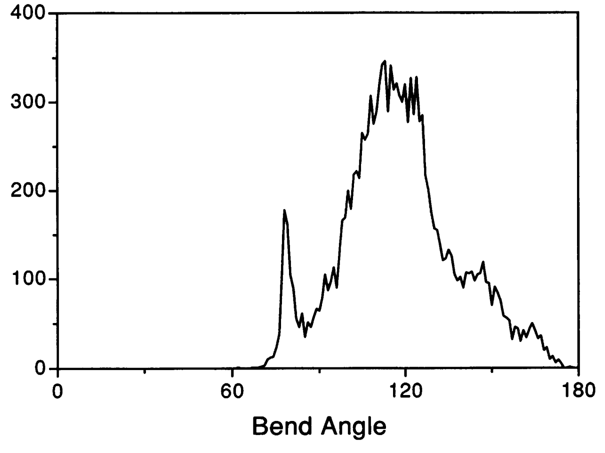


FIGURE 9 $C_{\alpha i}$–$C_{\alpha i+1}$–C_{Mi+1} pseudo-bond bend angle distribution (C_M is the center of steric mass). The counts in the database are plotted in one-degree bins.

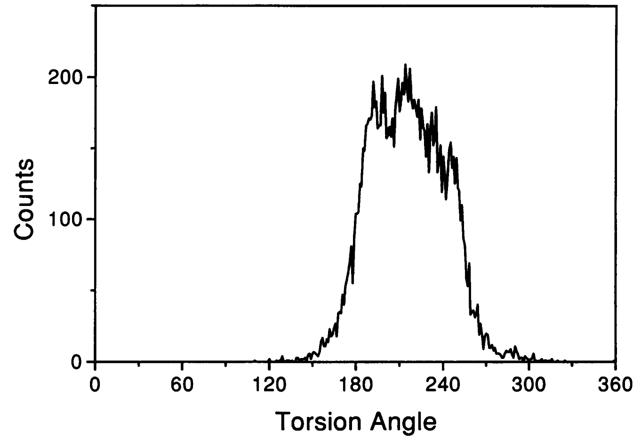


FIGURE 10 Distribution of $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–C_{Mi+2} pseudo-bond torsion angle offset with respect to the $C_{\alpha i}$–$C_{\alpha i+1}$–$C_{\alpha i+2}$–$C_{\alpha i+3}$ torsion angle (C_M is the center of steric mass). The counts in the database are plotted in one-degree bins.

By taking into account the distributions of the center of steric mass bend and torsion angles with respect to the main chain, one could represent the polypeptide chain as a main chain made by C_α's with attached ellipsoids, whose centers of steric mass are located at a fixed length from the C_α, depending on the residue, and in a direction defined by the average bend angle and torsion angle of the C_α–C_β (C_α center of steric mass) pseudo-bond, with the rotation axis along the C_α–C_β (C_α center of steric mass) pseudo-bond.

We did not elaborate further on this model, because it could be better adjusted, depending on specific practical applications. Rather we investigated another necessary prerequisite for putting ellipsoids into use, i.e., the ellipsoid-ellipsoid interaction potential function. We have tested the two approximate force fields that were described in Methods on the set of five high-resolution protein structures and on the homeodomain of thyroid transcription factor 1.

The results that were obtained with the two force fields are very similar. Both the scaling factors k are very close to the corresponding values for spherical interacting atoms ($k = 0.81$ versus 0.71 for the "overlap model," and $k = 1.12$ versus 1 for the "repulsive core model"), i.e., in both models the linear dimensions of the calculated ellipsoids have to be scaled down by approximately 10%, which is not a large amount, to avoid positive energies. Moreover, all of the considered molecules exhibit energy minima located even closer the corresponding spherical symmetry k values. These results point to the conclusion that ellipsoids are able to faithfully reproduce the shapes of amino acids. A somewhat less satisfactory conclusion is reached when the energies that are computed by AMBER force field and the two ellipsoid models are compared (Fig. 11). The value of the energy parameter ϵ_{ij} is, in both models, larger than expected. The explanation for this is that most probably the

TABLE 3 Statistics of the ellipsoids representing side chains

Residue	I_p (Å)	I_t (Å)	ϑ_{CB} (degrees)	ϑ_{CM} (degrees)	Counts
ALA	2.55 (0.02)	1.90 (0.00)	0.0 (0.0)	0.0 (0.0)	1355
CYS	3.15 (0.05)	2.01 (0.01)	14.8 (1.4)	38.0 (2.7)	329
ASP	3.23 (0.05)	2.05 (0.02)	10.5 (2.1)	38.4 (3.3)	945
GLU	3.88 (0.22)	2.09 (0.07)	10.2 (8.5)	37.0 (18.2)	872
PHE	4.18 (0.05)	2.40 (0.01)	13.4 (1.6)	55.4 (4.2)	610
HIS	3.96 (0.06)	2.19 (0.02)	14.3 (2.1)	53.6 (4.2)	370
ILE	3.46 (0.11)	2.33 (0.04)	32.4 (12.8)	44.6 (16.2)	815
LYS	4.54 (0.27)	2.09 (0.09)	10.8 (7.1)	38.8 (17.7)	1013
LEU	3.48 (0.06)	2.31 (0.02)	3.7 (2.7)	30.0 (4.5)	1265
MET	4.14 (0.29)	2.20 (0.09)	13.0 (8.5)	37.6 (21.0)	302
ASN	3.25 (0.06)	2.07 (0.02)	10.8 (2.6)	39.2 (4.0)	691
PRO	2.94 (0.02)	2.22 (0.01)	30.8 (1.7)	72.0 (2.5)	703
GLN	3.89 (0.23)	2.10 (0.08)	10.1 (9.0)	37.1 (18.9)	511
ARG	5.12 (0.35)	2.21 (0.11)	13.5 (6.5)	39.5 (18.8)	638
SER	2.70 (0.03)	1.93 (0.01)	9.8 (0.5)	22.7 (1.0)	1121
THR	2.87 (0.04)	2.15 (0.01)	20.5 (1.6)	33.6 (1.9)	931
VAL	2.79 (0.03)	2.37 (0.01)	30.3 (23.5)	35.8 (20.0)	1100
TRP	4.66 (0.12)	2.59 (0.05)	20.2 (6.4)	59.8 (16.9)	228
TYR	4.43 (0.05)	2.36 (0.01)	14.0 (1.6)	57.7 (4.2)	593
ALA	1.90 (0.00)	1.90 (0.00)	—	—	1355
CYS	2.76 (0.02)	1.88 (0.00)	43.9 (3.0)	67.0 (3.9)	329
ASP	2.66 (0.02)	2.00 (0.01)	39.5 (2.7)	67.6 (4.1)	945
GLU	3.23 (0.06)	2.05 (0.02)	22.0 (15.6)	47.7 (24.1)	872
PHE	3.60 (0.02)	2.39 (0.01)	24.5 (1.7)	66.5 (4.4)	610
HIS	3.39 (0.03)	2.14 (0.01)	28.8 (3.4)	68.0 (5.3)	370
ILE	3.34 (0.13)	2.15 (0.04)	57.8 (7.7)	67.8 (9.6)	815
LYS	3.91 (0.21)	2.05 (0.07)	19.9 (11.2)	46.1 (20.7)	1013
LEU	2.77 (0.02)	2.37 (0.01)	52.0 (23.5)	56.6 (21.9)	1265
MET	3.57 (0.21)	2.16 (0.07)	26.1 (13.1)	46.4 (24.4)	302
ASN	2.68 (0.02)	2.03 (0.01)	37.8 (4.5)	66.4 (6.0)	691
PRO	2.91 (0.03)	2.02 (0.01)	66.9 (1.0)	72.2 (1.5)	703
GLN	3.24 (0.07)	2.08 (0.02)	21.3 (16.2)	47.0 (24.4)	511
ARG	4.51 (0.28)	2.14 (0.09)	22.8 (9.6)	47.1 (19.5)	638
SER	2.28 (0.02)	1.77 (0.00)	57.6 (3.9)	70.5 (4.3)	1121
THR	2.72 (0.03)	1.93 (0.01)	69.0 (2.2)	79.8 (2.8)	931
VAL	2.97 (0.04)	2.01 (0.01)	88.6 (1.2)	88.2 (1.6)	1100
TRP	4.23 (0.03)	2.52 (0.01)	32.7 (9.3)	67.2 (21.0)	228
TYR	3.87 (0.02)	2.34 (0.01)	23.6 (1.8)	67.3 (4.5)	593

Major (I_p) and minor (I_t) axis values for ellipsoids representing amino acid side chains. The values were calculated alternatively by including (upper) or not (lower) the C_α carbon. The angles between the rotation (major) axis and the vector joining C_α to the center of steric mass is given as ϑ_{CM}, and the angle between the rotation (major) axis and the bond C_α–C_β is given as ϑ_{CB}. Averages are given with root mean square deviations in parentheses. Counts for each residue are listed in the last column.

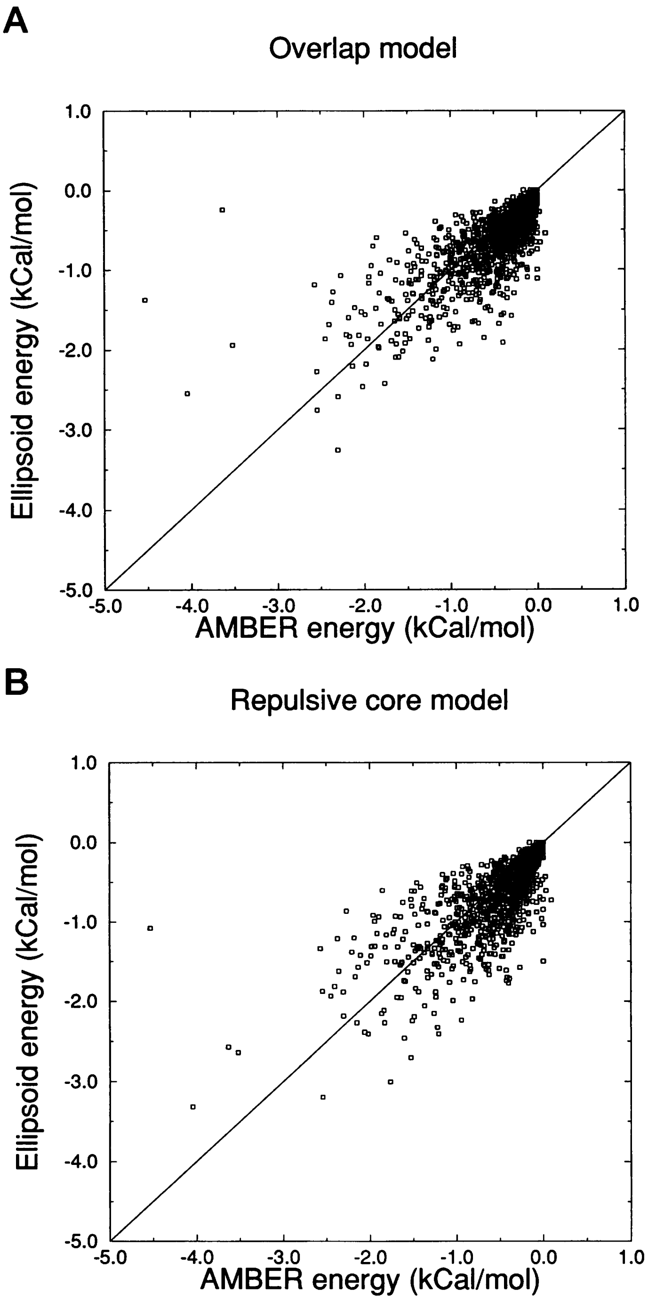


FIGURE 11 Residue-residue energy computed using ellipsoids versus all-atom AMBER van der Waals energy. The ellipsoid energy is computed using the "overlap model" (*a*) and the "repulsive core model" (*b*).

ellipsoid Lennard-Jones potential cannot properly fit large contact energies, as can be expected because, for short distances, a sum of 6–12 potentials is poorly reproduced by a single 6–12 potential, even though it is asymmetrical. To obtain a reasonable fit to the energy values, the energy constant ϵ_{ij} must be overestimated, and as a consequence, small energy values, which correspond to larger distances, appear to be constantly overestimated.

Notwithstanding these problems, it is remarkable that the correlation coefficient between AMBER and ellipsoid energies obtained with both models is 0.87, in view of the drastic simplification adopted.

Surprisingly, a test performed using a sphere instead of an ellipsoid representation of side chains (employing volume parameters derived from the ellipsoid model) leads to an even slightly better correlation coefficient (0.90). However, contrary to the ellipsoid models, energy minima for the five proteins are obtained for values of the scaling factor *k* consistently (20–30%) lower than the actual one (1.07). This means that, although the correlation coefficient is high, the intersphere distances are (globally) substantially larger than the Lennard-Jones optimal ones. This fact may lead to structure distortions upon energy minimization. It is worth mentioning, however, that this result as a sphere model might be easily incorporated into most available molecular modeling software packages.

A test was also performed on the homeodomain of thyroid transcription factor 1, a small molecule stabilized by a densely packed hydrophobic core, and by hydrophobic interactions among close residues in helices. Although the ellipsoid and AMBER energy values cannot be safely compared, because of the extensive (2000 steps) energy minimization that was performed and that consistently lowered the energy compared to the five-database high-resolution structures (see Methods), it is interesting that the largest AMBER residue-residue energies are also the largest residue-residue energies found with both ellipsoid potential models. Furthermore, for aromatic residues, for which modeling as an ellipsoid may appear to be too rough an approximation, the model does not break down. On the contrary, the large contact energy between Trp 48 and His 52, which is the largest calculated with the AMBER force field, is the second and the third largest energy in the "overlap" and "repulsive core" models, respectively, and the contact between Phe 20 and Phe 49, which is the fifth largest AMBER energy, is within the largest 25 values in both models. A test was also performed to check whether the ellipsoid representation is able to detect large contact AMBER energies and, conversely, whether large elliposid energies actually correspond to large AMBER energies. After discarding the few positive energies (three in total) in the simulations due to the extensive minimization and to constraints on the structure, only 10 (13) of the largest 100 AMBER energies were below 0.1 kcal/mol in the ellipsoid simulation with the "overlap model" ("repulsive core model"). The reverse test shows that all of the 100 largest ellipsoid contact energies that were computed in the ellipsoid simulation with the "overlap model" ("repulsive core model") are above 0.1 kcal/mol (0.4 kcal/mol) in the AMBER simulation.

CONCLUSIONS

We have analyzed the statistical behavior of proteins when considered as simplified representations. The reported results may help in the choice or design of appropriate models of protein structures for all those purposes that cannot be pursued by means of an all-atom representation. Our suggestion is that one may fix the bond lengths and bend angles in a pseudo-bond model of proteins and maintain the C_α trace pseudo-dihedrals as the only degrees of freedom. A promising idea is that of exploiting ellipsoids to model side-chain elongated shapes, although other degrees of freedom appear necessary to properly orient the ellipsoids. A better degree of accuracy in all of these representations could be attained by tuning rigid parameters to single amino acids or amino acid types or taking into account the correlation between different structural quantities, similar to previous approaches for the representation of proteins as C_α chains or C_α chains with attached C_β's (De Witte and Shakhnovich, 1994; Rey and Skolnick, 1992; Oldfield and Hubbard, 1994).

We wish to thank Dr. J. Carver (University of Wollngong, Australia) for reading the manuscript and making useful comments on both the form and content.

REFERENCES

Abagyan, R., and M. Totrov. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235:983–1002.

Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry.* 33:10026–10036.

Allen, M. P., and D. J. Tildesley. 1987. Computer Simulations of Liquids. Oxford University Press, Oxford.

Berne, B. J., and P. Pechukas. 1972. Gaussian model potentials molecular interactions. *J. Chem. Phys.* 56:4213–4216.

Brower, R. C., G. Vasmatzis, M. Silverman, and C. Delisi. 1993. Exhaustive conformational search and simulated annealing for models of lattice peptides. *Biopolymers.* 33:329–334.

Cattarinussi, S. 1994. Epigen 2.0, Manuale utente. Italfarmaco, Milan, Italy.

Cattarinussi, S., and G. Jug. 1990. Coil-globule transition temperature enhancement in a polymer molecule adsorbed to a wall. *J. Phys. A.* 23:2701–2706.

Cattarinussi, S., and G. Jug. 1991. Self-avoiding self-attracting model of polymer collapse and absorption. *J. Phys. II.* 1:397–419.

De Witte, R. S., and E. I. Shakhnovich. 1994. Pseudodihedrals: simplified protein backbone representation with knowledge based energy. *Protein Sci.* 3:1570–1581.

Eisenmenger, F., P. Argos, and R. Abagyan. 1993. A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* 231:849–860.

Fogolari, F., G. Esposito, P. Viglino, G. Damante, and A. Pastore. 1993. Homology model building of thyroid transcription factor 1 homeodomain (TTF-1 HD). *Protein Eng.* 6:513–520.

Gay, J. G., and B. J. Berne. 1981. Modification of the overlap potential to mimic a linear site-site potential. *J. Chem. Phys.* 74:3316–3319.

Gerber, P. R. 1992. Peptide mechanics: a force field for peptides and proteins working with entire residues as smallest units. *Biopolymers.* 32:1003–1017.

Godzik, A., A. Kolinski, and J. Skolnick. 1993. Lattice representation of globular proteins: how good are they? *J. Comp. Chem.* 14:1194–1202.

Goldstein, H. 1965. Classical Mechanics. Addison-Wesley Publishing, Reading, MA.

Goldstein, R. A., Z. A. Luthey-Schulten, and P. G. Wolynes. 1992. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Nat. Acad. Sci. USA.* 89:9029–9030.

Hinds, D. A., and M. Levitt. 1992. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA.* 89:2536–2540.

Hinds, D. A., and M. Levitt. 1994. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* 243:668–682.

Hunt, N. G., L. M. Gregoret, and F. E. Cohen. 1994. The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. *J. Mol. Biol.* 241:214–225.

Kolinski, A., and J. Skolnick. 1994a. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct. Funct. Genet.* 18:338–352.

Kolinski, A., and J. Skolnick. 1994b. Monte Carlo simulations of protein folding. II. Application to protein A, ROP and Crambin. *Proteins Struct. Funct. Genet.* 18:353–366.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.

Levitt, M. and A. Warshel. 1975. Computer simulation of protein folding. *Nature.* 253:694–698.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.

Oldfield, T. J., and R. E. Hubbard. 1994. Analysis of C_α geometry in protein structures. *Proteins Struct. Funct. Genet.* 18:324–337.

Orengo, C. A., T. P. Flores, W. R. Taylor, and J. M. Thornton. 1993. Identification and classification of protein fold families. *Protein Eng.* 6:485–500.

Pastore, A., R. A. Atkinson, V. Saudek, and R. J. P. Williams. 1991. Topological mirror images in protein structure computation: an underestimated problem. *Proteins Struct. Funct. Genet.* 10:22–32.

Rackovski, S. 1990. Quantitative organization of the known protein X-ray structures. I. Methods and short length scale results. *Proteins Struct. Funct. Genet.* 7:378–402.

Rey, A., and J. Skolnick. 1992. Efficient algorithm for the reconstruction of a protein backbone from the C_α-carbon coordinates. *J. Comp. Chem.* 13:443–456.

Šali, A., E. I. Shakhnovich, and M. Karplus 1994. How does a protein fold? *Nature.* 369:248–251.

Scheraga, H. A. 1993. Searching conformational space. *In* Computer Simulations of Biomolecular Systems. Theoretical and Experimental Applications. W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors. Escom Science Publishers B. V., Leiden, The Netherlands.

Sedgewick, R. 1990. Algorithms in C. Addison-Wesley Publishing, Reading, MA.

Sikorski, A., and J. Skolnick. 1990. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. II. α-Helical motifs. *J. Mol. Biol.* 212:819–836.

Skolnick, J., and A. Kolinski. 1990a. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key β-barrel proteins. *J. Mol. Biol.* 212:787–817.

Skolnick, J., and A. Kolinski. 1990b. Simulation of the folding of a globular protein. *Science.* 250:1121–1125.

Van Gunsteren, W. F., and H. J. C. Berendsen. 1987. GROMOS Manual. Biomos, Groningen, The Netherlands.

Verlet, L. 1967. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* 159: 98–103.

Viglino, P., F. Fogolari, S. Formisano, N. Bortolotti, G. Damante, R. Di Lauro, and G. Esposito. 1993. Structural study of rat thyroid transcription factor 1 homeodomain (TTF-1 HD) by nuclear magnetic resonance. *FEBS Lett.* 336:397–402.

Weiner, S. J., P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, S. Profeta, Jr., and P. Weiner. 1984. A new forcefield for mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106:765–784.

Weiner, S. J., P. A. Kollman, D. T. Nguyen, and D. A. Case. 1986. An all atom forcefield for simulations of proteins and nucleic acids. *J. Comp. Chem.* 7:230–252.

Williams, M. A., J. M. Goodfellow, and J. M. Thornton. 1994. Buried water and internal cavities. *Protein Sci.* 3:1224–1235.

Wolynes, P. G., J. N. Onuchic, and D. Thirumalai. 1995. Navigating the folding routes. *Science.* 267:1619–1620.